GENETIC ALGORITHM OPTIMIZED K NEAREST NEIGHBOR CLASSIFICATION FRAMEWORK (gaKnn)

D. G. N. Dayaratne

This dissertation was submitted to the Department of Computer Science and Engineering of the University Of Moratuwa in partial fulfillment of the requirement'! for the Degree of MSc in Computer Science

Department of Computer Science and Engineering
University of Moratuwa
Srilanka

November 2008

Abstract

Data classification attempts to assign a category or a class label to an unknown data object based on an available similar data set with class labels already assigned. K nearest neighbor (KNN) is a widely used classification technique in data mining. KNN assigns the majority class label of its closest neighbours to an unknown object, when classifying an unknown object. The computational efficiency and accuracy of KNN depends largely on the techniques used to identify its K nearest neighbours. The selection of a similarity metric to identify the neighbours and the selection of the optimum K as the number of neighbours can be considered as an optimization problem. The optimizing parameters for KNN are value for K. weight vector, voting power of neighbours, attribute selection and instance selection. Finding these values is a search problem with a large search space .Genetic Algorithms (OA) are considered to provide optimum solutions for search problems with a large search space. The search space is defined by the application do main. There are multiple real world classification applications that can utilize a parameter optimized KNN. Due to this, there is various research work carried out on using Genetic Algorithms for optimizing KNN classification.

Even though multiple instances of research had been carried out on using GA to optimize KNN there is no software framework available, which could be easily adapted to various application domains. This research is aimed towards building a framework to carry out the optimization of KNN classification with the help of a Genetic Algorithm. The work includes identifying issues and best practices on designing a suitable framework. The developed framework provides a basic backbone for GA optimization of KNN while providing sufficient flexibility for the user, to extend it to specific application domains.

This work discusses the design and implementation of a minimalist gaKnn framework. It is expected that this would serve as a basis for future enhancements.